Construction of Objective Tests
Professors Marjorie Devine & Nevart Yaghlian

List of Steps

1. Planning for the test:

--Outline subject-matter content to be considered as the basis for the test.
--Identify learning outcomes to be measured by the test.
--Prepare table of specifications.
--Choose appropriate type(s) of test items for evaluation of learning outcomes as summarized in the table of specifications.

2. Preparing the test:

--Write test items according to rules of construction for the type(s) chosen.
--Select the items to be included in the test according to table of specifications.
--Review and edit items according to guidelines.
--Arrange items: decide on a) grouping of items, b) sequence of items within groups, c) sequence of groupings.
--Prepare directions for the test; if necessary, prepare directions for individual items (e.g., matching type) or for sections (e.g., negative form of one-best response type).
--Decide on method of scoring.

3. Analyzing and revising the test:

--Perform test analysis to determine difficulty, discrimination and reliability.
--Retain, edit as necessary, or discard items on basis of analysis outcomes.
--Revise the test as a whole if necessary.

TEST PREPARATION

The initial step in testing must be the determination of what you are attempting to measure. When this decision is made and only after it is made can you decide what type of test to use.

A. Rules for Construction Multiple Choice Items

1. Design each item to measure an important learning outcome.
2. Present a single, clearly formulated problem in the stem of the item.
3. State the stem of the item in simple, clear language.
4. Put as much of the wording as possible in the stem of the item.
5. State the stem of the item in positive form, wherever possible.
6. Emphasize negative wording whenever it is used in the stem of an item.
7. Make certain that the intended answer is correct or clearly best.
8. Make all alternatives grammatically consistent with the stem of the item and parallel in form.
9. Avoid verbal clues which might enable students to select the correct answer or to eliminate an incorrect alternative:

--Similarity of wording in both the stem and the correct answer is one of the more obvious clues. Key words in the stem may be unintentionally repeated verbatim in the correct answer, a synonym may be used, or the words may simply sound or look alike.

--Stating the correct answer in textbook language or stereotyped phraseology may cause the student to select it because it looks better than the other alternatives, or because he vaguely recalls having seen it before.

--Stating the correct answer in greater detail may provide a clue. Also, when the answer is qualified by modifiers which are typically associated with the true statements (e.g., sometimes, may, usually), it is more likely to be chosen.

--Including absolute terms in the distracters enables students to eliminate them as possible answers, because such terms are commonly associated with false statements (e.g., always, never, all, none, only). This makes the correct answer obvious or at least increases the chances of guessing it.

--Including two responses that are all-inclusive makes it p possible to eliminate the other alternatives, since one of the two obviously must be the correct answer.

--Including two responses that have the same meaning makes it possible to eliminate them as potential answers. If two alternatives have the same meaning and only one answer is to be selected, it is fairly obvious that both of them must be incorrect.

10. Make the distracters plausible and attractive to the uninformed

    --Use the common misconceptions, or common errors, of students as distracters.
    --State the alternatives in the language of the student.
    --Use good sounding words (e.g., accurate, important, etc.) in the distracters, as well as in the correct answers.
    --Make the distracters similar to the correct answer in both length and complexity of wording.

11. Vary the relative length of the correct answer to eliminate length as a clue.
12. Avoid use of the alternative "all of the above" and use "none of the above" with caution.
13. Vary the position of the correct answer in a random manner.
14. Control the difficulty of the item either by varying the problem in the stem or by changing the alternatives.
15. Make certain each item is independent of the other items in the test.
16. Use an efficient time format.

Matching Items

1. Include only homogeneous material in each matching item.
2. Keep the lists of items short and put the brief responses on the right.
3. Use a larger, or smaller, number of responses than premises, and permit the responses to be used more than once.
4. Specify in the directions the basis for matching and indicate that each response may be used once, more than once, or not at all.

True-False Items

1. Include only one central significant idea in each statement.
2. Word the statement so precisely that it can be judged unequivocally true or false.
3. Use negative statements sparingly and avoid double negatives.
4. Statements of opinion should be attributed to some source.
5. Keep statements short and use simple language structure.

B. Guidelines for Preparing the Final Draft Reviewing, Selecting and Editing Items

1. Does each test item measure an important learning-outcome included in the table of specifications?
2. Is each item type appropriate for the particular learning outcome to be measured?
3. Does each item present a clearly formulated task?
4. Is the item stated in simple, clear language?
S. Is the item free from extraneous clues?
6. Is the difficulty of the item appropriate for the students to be tested?
7. Is each test item independent and are the items, as a group, free from overlapping?
8. Do the items to be included in the test provide adequate coverage of the table of specifications?

Arranging Items

1. The items should be arranged so that all items of the same type are grouped together.
2. The items should be arranged in order of increasing difficulty.
3. For some purposes, it may be desirable to group together items which measure the same learning outcomes or the same subject-matter content.

Writing Directions

The directions for an achievement test should be simple and concise and yet contain information concerning each of the following:

1. Purpose of the test.
2. Time allowed to complete the test.
3. How to record the answers.
4. Whether to guess when in doubt about the answer.

C. Other Considerations

Reproducing the Test

Decisions need to be made regarding page size, type size, page layout (length of line, placing of items on page, provision for response, page numbers, arrangement of alternatives in multiple choice items), preparation of copy, proofreading.

Administering and Scoring the Test

Decisions need to be made regarding setting of time limits, observation of time limits, physical set-up, proctors distribution of test, scoring sheets, scoring method (by hand/machine).

Assigning Grades

Decisions need to be made regarding weighting of items according to table of specifications.

Reporting Grades to Students

Plans need to be made for each of the following alternatives: posting by name, posting by number or returning graded tests to students.

Ground Rules Regarding Student "Complaints"

Decisions need to be made regarding procedure for appeal by students as well as method of dealing with appeals.

ANALYSIS OF TESTING

Item Analysis--The systematic evaluation of the effectiveness of each item of a test.

An item analysis can tell us:

1. The difficulty of the item.
2. The discriminating power of the item.
3. The effectiveness of each alternative.

Simplified Item-Analysis Procedure

There are a number of different item-analysis procedures that might be applied (Downie, 1967). For informal achievement tests used in teaching, only the simplest of procedures seems warranted. The following steps outline a simple but effective procedure. We shall use 32 test papers to illustrate the steps.

1. Arrange the test papers (all 32 papers) in order from the highest score to the lowest score.

2. Select approximately one-third of the papers with the highest scores and call this the upper group (10 papers). Select the same number of papers with the lowest scores and call this the lower group (10 papers). Set the middle group of papers aside (12 papers). Although these could be included in the analysis, using only the upper and lower groups simplifies the procedure.

3. For each item, count the number of students in the upper groups who selected each alternative., Make the same count for the lower group.

4. Record the count, from step 3, on a copy of the test, in columns to the left of the alternatives to which each count refers. The count may also be recorded on the item card or on a separate sheet, as follows:

item 1.

Alternatives	A	B*	C	D	E
Upper 10	0	6	3	1	0
Lower 10	3	2	2	3	0

* = correct answer

5. Estimate item difficulty, by determining the percentage of students who got the item right. The simplest procedure is to base this estimate on only those students included in the item-analysis groups. Thus, sum the total number in the upper and lower groups (10 + 10 = 20); sum the number selecting the correct answer. (For item 1. above, 6 + 2 = 8); divide the first sum into the second and multiply by 100, as follows:

$$\text{Index of Difficulty} = \frac{8}{20} \times 100 = 40\%$$

Although our computation is based on the upper and lower groups only, this provides a close approximation to the estimate that would be obtained with the total group. Thus, it is proper to say that the index of difficulty for this item is 40 per cent (for this particular group). Note that, since difficulty refers to the percentage getting the item right, the smaller the percentage figure the more difficult the item.

6. Estimate item discriminating power, by comparing the number of students in the upper and lower groups who got the item right. Note in our sample item above, that six students in the upper group and two students in the lower group selected the correct answer. This indicates positive discrimination, since the

item differentiates between students in the same way that the total test score does. That is, students with high scores on the test (upper group) got the item right more frequently than students with low scores on the test (lower group).

Although analysis by inspection may be all that is necessary for most purposes, an index of discrimination can be easily computed. Simply subtract the number in the lower group who got the item right from the number in the upper group who got the item right and divide by the number in each group. Thus, for our sample item, the computation would be as follows:

$$\text{Index of Discrimination} = \frac{6 - 2}{10} = .40$$

Note that the discriminating power of an item is reported as a decimal fraction. Maximum positive discriminating power is indicated by an index of 1.00. This is obtained only when all students in the upper group select the correct answer and no one in the lower group does. For our illustrative upper and lower groups of 10, the computation for an item with maximum discriminating power would be as follows:

$$\text{Index of Discrimination} \quad \frac{10 - 0}{10} = 1.00$$

Note that this item is at the 50 per cent level of difficulty (upper 10 got it right, lower 10 missed it). This explains why test makers are encouraged to prepare items at the 50 per cent level of difficulty. It is only at this level that maximum discrimination is possible.

Zero discriminating power (.00) is obtained when an equal number of students in both groups get the item right, and negative discriminating power when more students in the lower group than in the upper group get it right. Both types of items should be removed from general achievement tests and be discarded or improved.

7. Determine the effectiveness of the distracters, by comparing the number of students in the upper and lower groups who selected each incorrect alternative. A good distracter will attract more students from the lower group than the upper group. Thus, for our illustrative item-analysis data, in step 4, alternatives A and D are functioning effectively, alternative C is poor since it attracted more students from the upper group and alternative E is completely ineffective since it attracted no one. An analysis such as this is useful in evaluating a test item, and, when combined with an inspection of the item itself, it provides helpful information for improving the item.

The above procedural steps for analyzing items can be modified to fit particular situations. In some cases, inspecting the data, rather than computing the difficulty and discriminating power, may be all that is necessary. Also, in selecting the upper and lower groups, it may be desirable to use the top and bottom 25 per cent if the group is large, or the upper and lower halves if the group is small. The important thing is to use a large enough fraction of the group to provide useful information. Selecting the top and bottom 27 per cent of the group (as is recommended for more refined analysis) and applying other statistical refinements is seldom warranted with classroom achievement tests.

Simplified Methods of Treating Test Scores.

Test scores are normally described by two measures:

1. Average score or measure of central tendency.
2. Spread of scores or measure of variability.

Three types of averages:

1. Mean
2. Median
3. Mode

Two ways of describing variability:

1. Range
2. Standard deviation

Determining median and range.

See Chart #I

Determining the mean and standard deviation.

$$\text{mean} = \frac{\text{sum of all scores}}{\text{total number of students taking exam}}$$

$$\text{standard deviation} = \frac{\text{sum of high sixth} - \text{sum of low sixth}}{\text{half the number of students}}$$

Validity and Reliability

Valid tests measure what they actually were designed to measure.

Tests of validity:

1. Content
2 . Criterion - related
3. Construct

Reliable tests measure what they were designed to measure consistently.

Methods of determining reliability:

1. Test - retest method.
2. Equivalent - forms method.
3. Test - retest with equivalent forms.
4. Internal consistency method.

Chart I.

Frequency Distribution of Test Scores
for an Objective Test of 40 Items

test score	frequency	
38	*1*	
37	*1*	
36	*0*	
35	*2*	
34	*1*	
33	*2*	
32	*3*	
31	*2*	
30	*1*	
29	*4*	*- - - Median*
28	*2*	
27	*2*	
26	*2*	
25	*3*	
24	*1*	
23	*0*	
22	*1*	
21	*1*	
20	*0*	
19	*1*	
	N = 30	

OBJECTIVE TESTS
CONSTRUCTION OF MULTIPLE CHOICE ITEMS

I. Criteria for Multiple-Choice Questions [1]

"No research scientist would rely on a measurement without having been convinced about the accuracy of the measuring instrument. The same principle applies to educational measurement. A detailed analysis for each item in a test is basic to an understanding of the performance of students."
This analysis involves consideration of three issues:

A. The level of difficulty of an individual test item.

B. The extent to which the item succeeds in discriminating between the more knowledgeable and the less knowledgeable students.

C. The quality of the item in terms of its relevance to educational objectives of a particular course.

II. Types of Multiple-Choice Questions [1 and 3]

A. One-Best Response Type

This is the traditional and most frequently used type of multiple choice item. Each item consists of a stem (e.g., a statement, question, case history, situation, chart, graph or picture) followed by a series of four or five suggested answers for a question or completions for a statement. The suggested answers or completions other than the one correct (best) choice are called distractors.

In a five-choice item, a student knowing nothing about the subject matter has a one-out-of-five (20 per cent) chance of choosing the correct answer (best response) by random guessing.

EXAMPLES (One-Best Response Type):

1. The most effective prophylactic agent for the prevention of recurrences of rheumatic fever is:

   (a) acetylsalicylic acid
   (b) para-aminobenzoic acid
   (c) adrenocortiocotropic hormone
   (d) *penicillin
   (e) cortisone

In this example the stem might have been written as a question: "Which of the following is the most effective prophylactic agent for the prevention of recurrences of rheumatic fever?" There is little advantage in an incomplete statement over a question, but the incomplete statement is sometimes preferred because it can often be expressed in a simpler way with fewer words.

2. The above item could also be written as follows:

The most effective prophylactic agent for the prevention of recurrences of rheumatic fever is:

(a) acetylsalicylic acid
(b) para-aminobenzoic acid
(c) adrenocortiocotropic hormone
(d) cortisone
(e) *none of the above

In this manner the student must search his mind for the prophylactic agent most effective in the prevention of recurrence of rheumatic fever without having this agent, penicillin, suggested to him by finding it as one of the possible choices.
*Correct responses are identified by an asterisk.

3. Again, using the previous item as an example, it might be written as follows:
The most effective prophylactic agent for the prevention of recurrences of rheumatic fever is:

(a) acetylsalicilic acid
(b) para-aminobenzoic acid
(c) *penicillin
(d) cortisone
(e) none of the above

In this version, the "none of the above" response is incorrect and the correct response becomes c. It should be made clear to the student that the "none of the above" choice may sometimes be correct and sometimes incorrect.

4. Another variant of this type of item is the negative form: all but one of the choices are applicable and the student is asked to select the one which does not apply, or applies least, or is an exception in some way. The following is an item of this type:

Active immunization is available against all of the following diseases EXCEPT:

(a) tuberculosis
(b) smallpox
(c) poliomyelitis
(d) *malaria
(e) yellow fever

This "all of the following EXCEPT," however, requires a. switch from positive to negative thinking; this may throw the student off the track to the extent that an incorrect response might indicate a failure to follow the technique of the test rather than a true lack of knowledge of the subject. To avoid this possible difficulty, negative stems are usually placed together in a separate section of the test with special instructions calling attention to their negative form.

5. The best one-out-of-five type test item may follow, either singly), or in sets, the presentation of a case history or other situation presenting a problem that can be as complex as seems appropriate to the examiner. In presenting case histories or other problem situations, a clear and concise style is urged. The objective is to give the student all the necessary information-but only the information truly necessary to make the correct response out of the choices offered.

The following case history describes a patient with hemophilia, but the word "hemophilia" does not appear anywhere in the question. The student must, however, know about hemophilia in order to respond correctly to the three mutually independent questions that follow the case history.

A 14-year-old boy is admitted to the hospital with a nosebleed which followed slight trauma and which has persisted for four hours despite nasal packing. He has had repeated nosebleeds since early childhood. He had spontaneous hematuria on one occasion, and has ankylosis of both knees and the left elbow as a consequence of hemorrhage into these joints following injury. A maternal uncle was also a bleeder, but his mother, father, and two sisters have not had abnormal bleeding.

He has a blood-soaked pack in his nose and ankylosis of both knees and his left elbow. His spleen is not palpable. There is no evident lymphadenopathy; no petechiae or telangiectases are seen. The following laboratory data are reported: hemoglobin 13 gm per 100 ml; erythrocyte count 4,500,000 per cu mm; leukocyte count 12,000 per cu mm; differential count normal; platelets 460,000 per cu mm; urine shows no protein, red blood cells, or other abnormalities in the sediment.

1) Which of the following tests is most likely to show an abnormality?

(a) Tourniquet test
(b) Bleeding time
(c) *Clotting time
(d) Clot reaction
(e) Bone-marrow examination

2) An abnormality would be expected in

(a) one stage prothrombin time
(b) *thromboplastin consumption test
(c) plasma fibrinogen content
(d) platelet fragility
(e) none of the above

3) The most efficacious of the following therapeutic procedures would be:

(a) local applications of thrombin-soaked packs to the nose
(b) intravenous administration of a suspension of fresh, normal platelets in saline
(c) *intravenous administration of fresh plasma
(d) intravenous administration of fibrinogen and calcium gluconate
(e) intravenous administration of vitamin K

B. The Matching Type*

This type consists of two lists of terms, phrases, statements, portions of a diagram and directions for matching one of the responses on the second list to each of premises on the first list. (The distinction between premise and response is purely formal--premises are identified as bearing numbers and responses as bearing letters. Lists of four or five are preferred.)

EXAMPLES

1. Pick the phrase in column (B) that most accurately defines each of the phrases in column (A).

Column A	Column B
(B) 1. Long-standing silicosis	(A) Hypertrophy of the left ventrical

(E) 2. Constrictive pericarditis	(B) Cor pulmonale
(C) 3. Rheumatic heart disease	(C) Mitral and aortic stenosis
(A) 4. Systemic hypertension	(D) Subpulmonic stenosis
	(E) Congestive failure without cardiac enlargement

Care must be taken to be sure that the association keyed as the correct response is unquestionably correct and that the numbered item could not be rightly associated with any other choice.

2. In another form of matching item, the student is directed to select the A response if the word or phrase is associated with A only, B if the word or phrase is associated with B only, C if the word or phrase is associated with both A and B, or D if the word or phrase is associated with neither A nor B. The following is an example.

(A) Plasmodium vivax malaria
(B) Plasmodium falciparum malaria
(C) Both
(D) Neither

1. A combination of primaquine and chloroquine is the treatment of choice for an acute attack (A).

2. Clinical attacks are suppressed by ingestion of chloroquine once a week while in an endemic area (C).

3. Infection is prevented by ingestion of chloroquine once a week (D).

*Correct responses are indicated by the letter in parentheses on the numbered lists.

C. Multiple True-False Type

Multiple true-false items consist of a stem followed by four or five true or false statements. The stem may be in the form of a question, a statement, a case history, or clinical data presented in pictorial fashion.

EXAMPLE

In one form of this type of item, the student is directed to select A response if 1, 2 and 3 are correct; B if 1 and E are correct; C if 2 and 4 are correct; D if only 4 is correct; and E if all four are correct. In the following example (1) and (3) are correct; the correct response is therefore B.

A child suffering from an acute exacerbation of rheumatic fever usually has:

(1) an elevated erythrocyte sedimentation rate
(2) a prolonged P-R interval
(3) an elevated antistreptolysin 0 titer
(4) subcutaneous nodules

It should be noted that the code to this type of item is shown in abbreviated form at the top of every page where items of this type appear, as follows:

Directions Summarized

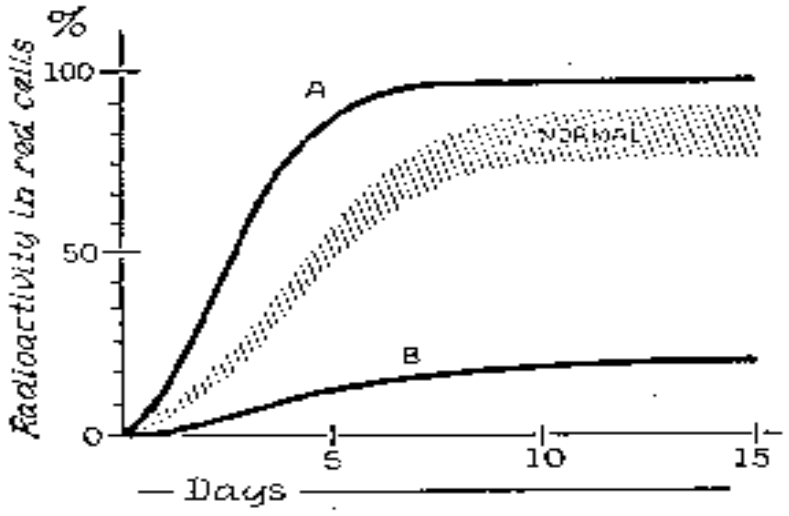
A	B	c	D	E
1, 2, 3 only	1, 3 only	2, 4 only	4 only	All are correct

D. Pictorial, Tabulated and Graphic Material

A single picture, graph or table may be presented and a set of questions developed to deal with it. Sets of two or more closely related pictures also may be used, followed by a series of questions requiring the students to utilize all of the visual material in choosing his/her answers.

Good use of questions based on a graph is illustrated by the following EXAMPLE:

The graph below shows incorporation of radioactive iron in erythrocytes of peripheral blood after an intravenous injection of radioactive iron citrate. Study this graph to answer



1. A 40-year-old man has had prolonged gastrointestinal bleeding. He has an iron uptake as shown in curve A. This observation implies all of the following information EXCEPT

   (A) increased erythrocyte production

(B) iron deficiency
(C) *saturated iron-binding protein
(D) reticulocytosis
(E) adequate serum folic acid concentration

2. A 40-year-old man has normal erythroid blood values (hematocrit 45 per cent, hemoglobin 14.6/100 ml. reticulocytes 1.0 per cent). His iron uptake is shown in curve B. These findings represent

(A) chronic glomerulonephritis
(B) *hemochromatosis
(C) agnogenic myeloid metaplasia
(D) congestive splenomegaly
(E) low serum transferrin concentration

Items accompanying graphs and charts may require the candidate to interpret the data and make certain deductions about them. When sets of illustrations are used and items call for a matching process, unintentional clueing is to be avoided.

III. COMPONENTS AND CHARACTERISTICS OF GOOD QUESTIONS[2] (Multiple-Choice form)

The Stem

a. should be in the form of a question or of a statement to be completed

b. should be expressed clearly and concisely, avoiding poor grammar, complex syntax, ambiguity and double negatives

c. should generally present a positive question. (If a negative is used it should be emphasized with italics or underlining.)

d. should generally ask for one answer only (the correct or the best answer)

e. should include as many as possible of the words common to all alternatives.

The Alternatives

a. each item should have either four or five alternatives, all of which should be mutually exclusive and not too long

b. all alternatives should follow grammatically from the stem and be parallel in grammatical form

c. the alternatives should be expressed simply enough to make clear the essential differences between them, and must be unambiguous

d. the intended answer should be clearly correct to the informed, while the distractors should be definitely incorrect, but plausible to the uninformed.

IV. Check List in Question Preparation[1]

To aid the teacher in maintaining acceptable standards in the preparation of multiple-choice questions, the following check list is offered. If an item does not satisfy each of these standards, it should be discarded or revised.

1. Does the item deal with one or more important aspects of the subject?

2. Does the item call for information that the student should know or be able to deduce without consulting a reference source?

3. Is the item appropriate for the level of knowledge expected of the student? Items that are too difficult or too easy cannot make effective discriminations among the students.

4. Is the central problem stated clearly and accurately? Wording that is ambiguous or fuzzy may mislead the student and destroy the validity of an item.

5. Have irrelevant clues to the correct response been avoided?

6. Is the item written with as few words as possible to make it clear and complete? Unnecessary words increase reading time; the test is intended to evaluate nutritional science knowledge, not reading speed.

7. Is the item type the best one for the particular point or problem?

8. Is the item written in conformity with the format?

References

1. From Hubbard, Evaluation in Medical Education

2. From Whitfield, R.C., Criteria of Quality for Multiple Choice Questions

3. From Lindquist, Educational Measurement

Examples of different varieties of multiple-choice items

A. The correct answer variety

Who invented the sewing machine?

a. Fulton
*b. Howe
c. Singer
d. White
e. Whitney

B. The best answer variety

What was the basic purpose of the Marshall Plan?

a. militarily defend Western Europe
*b. reestablish business and industry in Western Europe
c. settle United States differences with Russia
d. directly help the hungry and homeless in Europe

C. The multiple response variety

What factors are principally responsible for the clotting of blood?

a. contact of blood with a foreign substance
*b. contact of blood with injured tissue
c. oxidation of hemoglobin
*d. presence of unchanged pro thrombin

D. The incomplete statement variety

Millions of dollars worth of corn, oats, wheat, and rye are destroyed annually in the United States by

a. mildews
b. molds
c. rusts
*d. smuts

E. The negative variety

Which of these is NOT true of viruses?

a. Viruses live only in plants and animals.
b. Viruses reproduce themselves.
*c. Viruses are composed of very large living cells.
d. Viruses can cause diseases.

F The substitution variety

Passages to be read

Surely the forces of education should be fully utilized to acquaint youth with the real nature of the dangers to democracy, for no other place offers as good or better opportunities than the school for;

1 2

a rational consideration of the problems involved.

3

Items to be answered

1. *a. , for
   b. . For
   c. —for
   d. no punctuation needed
2. a. As good or better opportunities than
   b. as good opportunities or better than
   c. as good opportunities as or better than
   *d. better opportunities than
3. *a. rational
   b. radical
   c. reasonable
   d. realistic

The Incomplete alternatives variety*

An apple that has a sharp, pungent, but not disagreeably sour or. bitter taste is said to be (4).

a. p b. q *c. t d. v e. w

The combined response variety

In what order should these sentences be written in order to make a coherent paragraph?

A. A sharp distinction must be drawn between table manners and sporting manners.

B. This kind of handling of a spoon at the table, however, is likely to produce nothing more than an angry protest against squirting grapefruit juice about.

C. Thus, for example, a fly ball caught by an outfielder in baseball or a completed pass in football is a subject for applause.

D. Similarly, the dexterous handling of a spoon in golf to release a ball from a sand trap may win a championship match.

E. But a biscuit or a muffin tossed and caught at the table produces scorn and reproach.

a. A**.** B, C. D. E
*b**.** A, C. E, D. B
c. A, E. C. D, B
d. B, E, D. C. A

*The numeral in parentheses indicates the number of letters in the correct answers (which in this case is 'tart'). Using this convention serves to rule out borderline correct answers.

Adapted) from “Writing the Test Item” by R L. Ebel (1951) and A G Wesman (1971) in *Educational Measurement.* First and Seconf Editions; Reprinted with permission ot the American Council on Education.

Basic Principles of Test Construction

prepared by Damien Merchan
Distributed by the Office of Instructional Support
400 CCC; 255-3493

LIMITATIONS OF CONSTRUCTED-RESPONSE TESTS

Poor sampling of domain
Low inter-rater reliability
Low intra-rater reliability
Low construct validity
Poor discrimination within examinees
Time consuming to grade

Scores influenced by:

handwriting / layout / appearance
spelling
grammar
writing style, word choice, fluidity
expectation of the grader
length of essay
time of day (lower in evening)
responses of other people
presence / absence/ quality of scoring key

PRESENT A CLEAR PROBLEM IN THE STEM

POOR Adolescents obtain information about sex:

A. in order to reduce the possibility of pregnancy
B. from parents
C. from same-sex friends*
D. in a haphazard fashion

BETTER Most adolescents obtain information about sex from:

A. same sex friends*
B. parents
C. health service personnel
D. teachers

PUT ALTERNATIVES AT THE END OF THE QUESTION

POOR Before the Civil War, the South's:

A. emphasis on staple-crop production*
B. lack of a suitable supply of raw materials
C. short supply of personnel capable of operating the necessary machinery was one of the major reasons manufacturing developed more slowly than it did in the North.

BETTER Before the Civil War, manufacturing developed more slowly in the South than in the North. One of the major causes of this was the South's:

A. emphasis on staple-crop production*
B. lack of a suitable supply of raw materials
C. short supply of personnel capable of operating the necessary machinery

PUT MOST OF THE WORDING IN THE STEM

POOR In objective testing, the term objective:

A. refers to the method of identifying the learning outcomes
B. refers to the method of selecting the test content
C. refers to the method of presenting the problem
D. refers to the method of scoring the answers*

BETTER In objective testing, the term objective refers to the method of:

A. identifying the learning outcomes
B. selecting the test content
C. presenting the problem
D. scoring the answers*

AVOID UNNECESSARY WORDINESS

POOR For almost a century, the Rhine river has been used by Europeans for a variety of purposes. However, in recent years, the increased river traffic has resulted in increased levels of diesel pollution in the waterway. Which of the following would be the most dramatic result if, because of the pollution, the Council of Ministers of the European Community decided to close the Rhine to all shipping?

A. closure of the busy river Rhine ports of rotterdam, Marseilles and Genoa
B. increased prices for Ruhr products*
C. reduced competitiveness of the French Aerospace Industry
D. shortage of water for Italian industries

BETTER Which of the following would be the most dramatic result if, because of diesel pollution from ships, the river Rhine were closed to all shipping?

A. closure of the busy river Rhine ports of Rotterdam, Marseilles and Genoa
B. increased prices for Ruhr products*
C. reduced competitiveness of the French Aerospace Industry
D. shortage of water for Italian industries

AVOID NEGATIVELY WORDED STEMS

POOR Sometimes a teacher finds it necessary to use a mild form of punishment. when this occurs, which of the following should not happen?

A. Children should not believe all of their behavior is bad.
B. Children should understand the reason(s) why they are being punished.
C. Children should understand that the teacher, not the children, controls when the punishment will end.*

BETTER Sometimes a teacher finds it necessay to use a mild form of punishment. When this occurs, it is important that the children understand:

A. that it may be a long time before happy times return to the classroom

B. the reason(s) why they are being punished*
C. that the teacher, not the children, controls when the punishment will end

GOOD Thurstone's 7 primary mental abilities include all of the following EXCEPT:

A. word fluency
B. reasoning
C. social interaction*
D. perceptual speed

AVOID REQUIRING PERSONAL OPINION

POOR Which of the following men contributed most towards he defeat of Hitler's Germany in World War II?

A. Winston Churchill
B. Josef Stalin
C. Franklin D. Roosevelt
D. George Patton

AVOID LINKED OR CLUED ITEMS

PRECEDING ITEM

Question 1 The perimeter of a rectangle is 350 centimeters. The length of the rectangle is 3 centimeters longer than the width. What is the width?

A. 18.7
B. 86.0*
C. 89.0
D. 116.7

POOR SUBSEQUENT ITEM: LINKED TO QUESTION 1

Question 2. What is the area of the rectangle described in Question 1?

A. 1050 sq. cm.
B. 7396 sq. cm.
C. 7654 sq. cm.*
D. 8188 sq. cm.

BETTER SUBSEQUENT ITEM: INDEPENDENT OF QUESTION 1

Question 2. The width of a rectangle is 4 centimeters and the length is 3 centimeters. What is the area?

A. 9 sq. cm.
B. 12 sq. cm.*
C. 16 sq. cm.
D. 17 sq. cm.

ALL OPTIONS SHOULD BE HOMOGENEOUS

POOR What is the official state bird of Pennsylvania?

A. mountain laurel

B. Philadelphia
C. ruffed grouse*
D. ibex

BETTER What is the official state bird of Pennsylvania?

A. goldfinch
B. robin
C. ruffed grouse*
D. wild turkey

ALL OPTIONS SHOULD BE PLAUSIBLE

POOR Who succeeded Giscard d'Estaing as President of France in l981?

A. Georges Pompidou
B. Charles De Gaulle
C. Francois Mitterand*
D. Mick Jagger

PUT REPEATED WORDS IN THE STEM, NOT THE OPTIONS

POOR Which of the following is the best definition of a *seismograph* ?

A. An apparatus for measuring sound waves.
B. An apparatus for measuring heat waves.
C. An apparatus for measuring earthquake waves.*
D. An apparatus for measuring ocean waves.

BETTER I A *seismograph* is an apparatus for measuring:

A. earthquake waves*
B. heat waves
C. ocean waves
D. sound waves

BETTER II What type of waves does a *seismograph* measure?

A. earthquake*
B. heat
C. ocean
D. sound

MAKE ALL OPTIONS GRAMMATICALLY CONSISTENT WITH STEM

POOR: grammatical mistakes clues the answer.

An angle of 90^{o} is called a

A. acute angle
B. obtuse angles
C. right angle*

BETTER: all options have grammatical relation to the stem.

Angles of 90^{o} are called

A. acute angles
B. abtuse angles
C. right angles*

Green plants may lose their color when:

A. are forming flowers
B. grown in the dark*
C. are placed in strong light
D. temperature drops

Green plants may lost their color when:

A. flowers are formed
B. grown in the dark*
C. placed in strong light
D. surrounding temperatures drop

ORDER OPTIONS LOGICALLY

GOOD The average difference between males and females in the attainment of puberty is:

A. 1 year
B. 2 years*
C. 3 years
D. no difference

WORD LENGTH OF TRUE STATEMENTS = FALES STATEMENTS

T* F Sheldon's main point in his theory of body types was that, given a person's physical characteristics, it is possible to pridict his/her personality and behavior.

T F* The population of California is now 19 million.

RULES FOR WRITING MULTIPLE-CHOICE ITEMS

1. Design each item to measure an important learning outcome.
2. Present a single clearly formulated problem in the stem of the item.
3. Put the alternatives at the end of the question, not in the middle.
4. Put as much of the wording as possible in the stem.
5. Eliminate unnecessary wordiness
6. Avoid negatively worded stems. "Which of the following is not....."
7. Avoid requiring personal opinion. Other item types are more suitable for this.
8. Avoid textbook wording.
9. Do not have linked or clued items.
10. All options should be homogeneous.
11. All options should be plausible.
12. Put repeated words in the stem, not in the options
13. Punctuation should be consistent.
14. Make all options grammatically consistent with the stem of the item.
15. List options vertically.
16. Order options logically.
17. Use the option "all of the above" sparingly.
18. Use the option "none of the above" sparingly.

RULES FOR WRITING TRUE-FALSE ITEMS

1. Make sure that the item is definately true or false.
2. Avoid verbal clues (specific determiners) that indicate the answer.
3. Test important ideas rather than trivia.
4. Keep the word-length of true statements about the same as that of false statements.
5. Dont present items in easily learned pattern.
6. Don't copy sentences directly from textbooks and other written materials.

RULES FOR WRITING COMPLETION ITEMS

1. State the item so that only a single, brief answer is possible.
2. Start with a direct question and switch to an incomplete statement only when greater conciseness is possible by doing so.

   What is another name for cone-bearing trees? (coniferous trees)

   Cone-bearing trees are also called: (coniferous trees)
3. Leave only one blank. This should relate to the main point of the statement.
4. Place the blanks at the end of the statement.
5. Word the items to avoid irrelevant clues or specific determiners.

RULES FOR *CONSTRUCTING* ESSAY ITEMS

1. Define the behavior the student is expected to exhibit before writing the prompt.
2. Ask the student to use knowledge in novel situations rather than simply recalling information.
3. Ask questions that are relatively specific and focused and which will elicit relatively brief responses.
4. If you are using many essay questions in a test, ensure reasonable coverage of the course objectives.
5. Follow the test specifications in writing prompts. Questions should cover the subject areas as well as the complexity of behaviors cited in the test blueprint. Pitch the questions at the students' level.
6. Formulate questions that present a clear task to be performed.
7. Provide ample time for answering and suggest a time limit for each question.
8. Indicate the point value for each question.

9. Word questions calling for examinee opinion on controversial matters so that they ask the examinee to geve evidence to support the opinion and evaluate the examinee's response in terms of the evidence presented rather than the opinion or position taken.
10. Require all examinees to answer the same questions; don't give optional questions

RULES FOR *SCORING* ESSAY QUESTIONS.

1. Prepare some type of scoring guide (e.g., an outline, an "ideal" answer).
2. Score restricted-response answers by the point method, using a model answer as a guide.
3. Grade extended-response answers by the rating method, using defined criteria as a guide. Assign to 5 grade bands by general criteria. Then check all papers in each band for uniformity and make any necessary shifts.
4. Evaluate all of the students' answers to one question before proceeding to the next question.
5. Periodically rescore previously scored papers.
6. Evaluate answers to essay questions without knowing the identity of the writer.
7. Have two or more people grade each answer.

FRANCE TO REQUIRE VISAS FOR VISITORS TO COMBAT TERROR

NEW BLAST KILLS OFFICER

U.S. Travelers Among Those Needing a Permit Under Antiterrorist Measures

By RICHARD BERNSTEIN
Special to The New York Times

PARIS, Sept. 14 — Prime Minister Jacques Chirac, speaking minutes after a fatal bomb explosion in Paris, announced measures tonight to combat a wave of terrorism, including a requirement that Americans and most other foreign visitors obtain visas.

Reuters

Prime Minister Jacques Chirac yesterday announcing measures to fight terrorism.

Only citizens of the II other countries of the European Community and Switzerland will be exempt from the visa requirement, which Mr. Chirac said would take effect Monday. The move is part of a French effort to monitor better the movements of potential terrorists from abroad.

Policeman Is Killed

Mr. Chirac announced the measures minutes after a bomb explosion, the third here in a week, took place in a parking garage underneath a popular Parisian cafe, killing a policeman and seriously wounding two people.

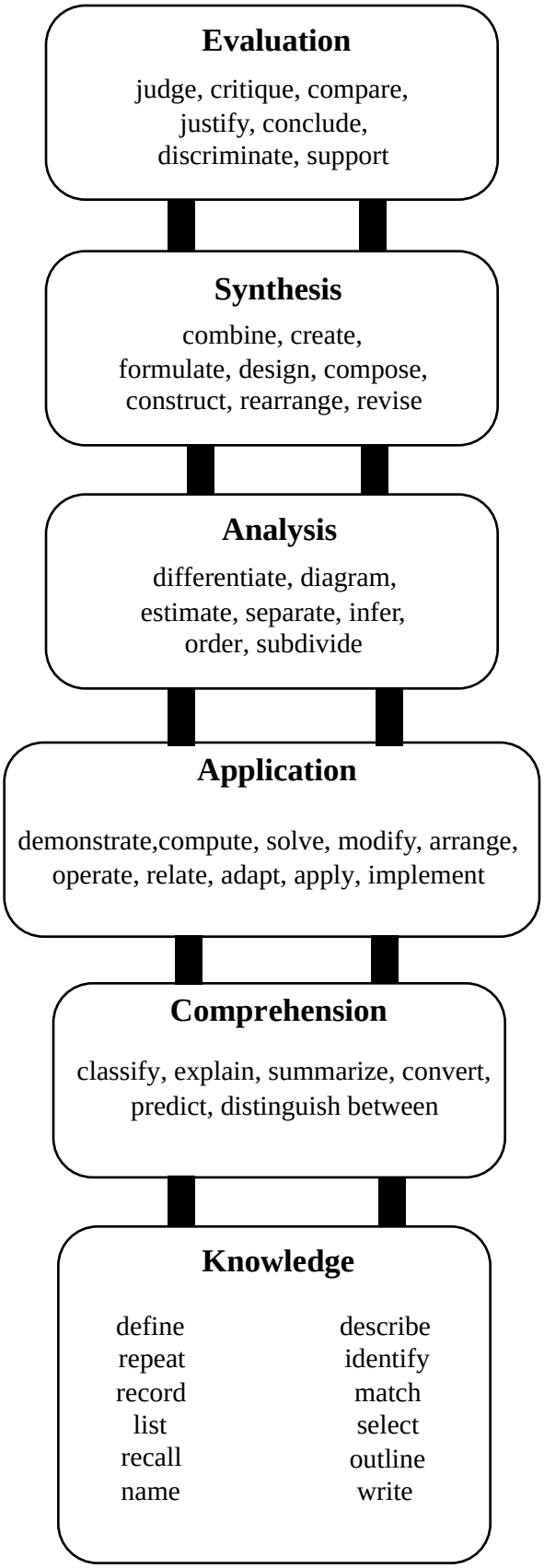
The attack was similar to eight others that have taken place here since December, provoking calls from politicians of the left and the right for strong measures to enhance security.

Vows 'Draconian' Measures

Speaking in a radio broadcast that had been scheduled before tonight's bombing, Mr. Chirac warned that France would take "draconian" measures against anybody it found to be backing the terrorist attacks.

"One of the people placing these bombs will inevitably be caught and he will talk," Mr. Chirac said. "And we will take draconian measures; we will be without pity for anybody, absolutely anybody found to be manipulating these terrorists.

"Terrorists should expect from the French Government no clemency of any sort, direct or indirect, official or unofficial, secret or open."


Evaluation
judge, critique, compare,
justify, conclude,
discriminate, support
Synthesis
combine, create,
formulate, design, compose,
construct, rearrange, revise
Analysis
differentiate, diagram,
estimate, separate, infer,
order, subdivide
Application
demonstrate,compute, solve, modify, arrange,
operate, relate, adapt, apply, implement
Comprehension
classify, explain, summarize, convert,
predict, distinguish between
Knowledge
define
repeat
record
list
recall
name
describe
identify
match
select
outline
write

Different Levels of Questioning:

Knowledge

What position does Mr. Chirac hold in France?

a) President
b) Prime Minister
c) General Secretary
Attorney General

Comprehension

How many countries are there in the European Community?

a) 10
b) 11
c) 12
d) 13

Application

It can be inferred from the passage the Mr. Chirac believes that the greatest impact of strict visa regulations is that they:

a) help the police to subdue French terrorist organizations
b) help the police to locate foreign terrorists operating in France
c) reduce the number of foreign terrorists entering France
d) give more power to anti-terrorist police

Analysis

Which of the following best explains why countries of the Eurpoean Community have been exempted from France's new visa regulations?

a) few terrorists operate in those countries
b) France is bound by its international treaties
c) those countries have signed the European Convention on the Suppression of Terrorism
d) France needs the support of those countries in its fight against terrorism

Synthesis

Briefly outline an alternative strategy that France could use in order to combat terrorist attacks. Your answer should include, a) a description of the problem, b) your proposal for dealing with it, and c) the likely consequences, both negative and positive, of such action. Your answer should not exceed 1 page. Suggested time for this question is 20 minutes. 10 points are allocated for this question.

Evaluation

Write a critique of the language used by Mr. Chirac, as reported in this article, in terms of its political flavor and intended audience. Support your discussion with relevant quotes from the passage in addition to examples drawn from other sources. Points will be awarded for

clarity of thinking, logical argumentation and relevant examples. Do not exceed 2 pages. 26 points are allocated for this question.

Additional Reading

- Gronlund, N. (1988). How to construct achievement tests. Engelwood Cliffs, NJ: Prentice-Hall.
- Nitko, A. (1983). Educational tests and measurement: An introduction. New York: Harcourt Brace Jovanovich.
- Heywood, J. (1977). Assessment in higher education. London: John Wiley & Sons.
- Ebel, R. (1972). Essentials of educational measurement. Englewood Cliffs, NJ: Prentice-Hall.
- Kirschenbaum, H., Simon, S. B. & Napier, R. W. (1971). Wad-Ja-Get? The grading game in American education. New York: Hart Publishing Company.

Weighting of Exams: Table of Specifications

Content Outline	Knowledge	Comprehension	Application	Analysis	Synthesis	Evaluation	Total
History of Adolescence	5	2	1				8
Physical Development	9	4	2	2		2	19
Psychological Implications	4	6		1	1	2	14
Cognitive Development	8	4	1	2	3	1	19
Total No. of Items	26	16	4	5	4	5	60